

# Automatic Visual Theme Discovery from Joint Image and Text Corpora

Ke Sun, Xianxu Hou, Qian Zhang, Guoping Qiu

School of Computer Science, University of Nottingham Ningbo China  
(Ke.Sun, Xianxu.Hou, Qian.Zhang, Guoping.Qiu)@nottingham.edu.cn

**Abstract.** A popular approach to semantic image understanding is to manually tag images with keywords and then learn a mapping from visual features to keywords. Manually tagging images is a subjective process and the same or very similar visual contents are often tagged with different keywords. Furthermore, not all tags have the same descriptive power for visual contents and large vocabulary available from natural language could result in a very diverse set of keywords. In this paper, we propose an unsupervised visual theme discovery framework as a better (more compact, efficient and effective) alternative to semantic representation of visual contents. We first show that tag based annotation lacks consistency and compactness for describing visually similar contents. We then learn the visual similarity between tags based on the visual features of the images containing the tags. At the same time, we use a natural language processing technique (word embedding) to measure the semantic similarity between tags. Finally, we cluster tags into visual themes based on their visual similarity and semantic similarity measures using a spectral clustering algorithm. We conduct user studies to evaluate the effectiveness and rationality of the visual themes discovered by our unsupervised algorithm and obtains promising result. We then design three common computer vision tasks, example based image search, keyword based image search and image labelling to explore potential application of our visual themes discovery framework. In experiments, visual themes significantly outperforms tags on semantic image understanding and achieve state-of-art performance in all three tasks. This again demonstrate the effectiveness and versatility of proposed framework.

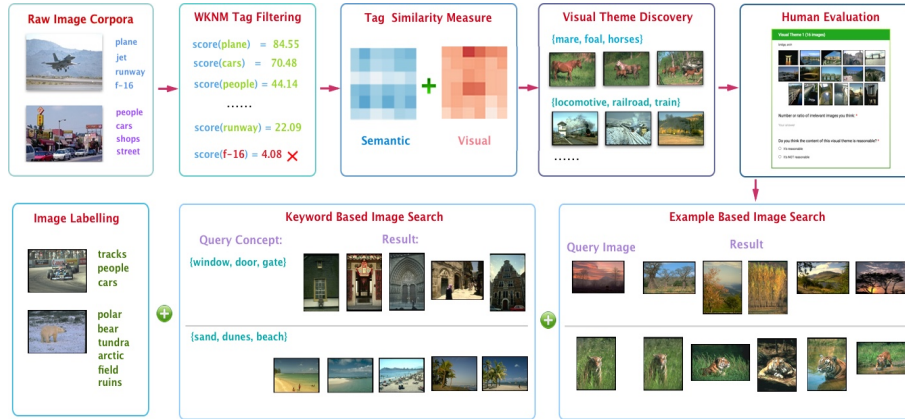
## 1 Introduction

The popularisation of photo sharing websites such as Flickr and Instagram encourages more and more people to share their life experience by uploading numerous images on a daily basis. Among various information contained in these images, associated tags are of great importance for helping computer vision (CV) algorithms to understand the semantic meaning of images. However, due to human’s subjectivity towards visual content understanding, different tags are often used to describe visually similar images. For example, images containing lakes, rivers and ocean could all be tagged with *water*, or we can just use *lake*, *river*,

*ocean* respectively. It's a very natural case but it often confuses CV algorithms since they are forced to distinguish similar visual instances.

Additionally, not all tags show strong connection to particular visual content, such as *wonderful* and *beautiful*, but they are frequently used in social media websites. Visually describing such kinds of tags is pretty challenging even for human themselves, let alone CV algorithms.

Another problem is the curse of dimensionality. Appearance of different tags in image annotations are often represented using one-hot encoding in order to be easily processed by CV algorithms. Hence, the dimensionality of annotation for a single image is the whole size of tag vocabulary. The overall annotation matrix would be extremely sparse since each image is only associated with a few tags. Traditional dimensionality reduction methods mainly focus on tag frequency, while the semantic and visual correlation between tags are often ignored.



**Fig. 1.** Overview of visual theme discovery framework and its applications. Given images and associated tags, we first eliminate less qualified tags using WKNM tag filtering method, then clustering tag into visual themes according to their semantic and visual similarities. Next we ask human evaluators to evaluate the quality of discovered visual themes. Applications of visual themes are shown at the bottom row.

To address the issues mentioned above, We propose to use Visual Theme (VT) as a replacement of tag-based annotation for compact visual content description. A visual theme, consisted of a small set of tags, is capable of describing a group of similar visual contents in images. Besides, tags within the same VT are also semantically related.

We develop a data-driven framework to automatically discover VTs from joint image and tag corpora. We start by examining each tag's ability for visual content description, then eliminate tags whose descriptive ability fall under certain level. Next we measure the pairwise semantic and visual similarity amongst

the remaining tags, then merge them into a joint similarity matrix. Visual similarity measures how tags are visually connected to describe visual contents, and semantic similarity measures how close tags are in natural language understanding. Finally we cluster tags into a collection of VTs according to the joint similarity matrix. The workflow of the proposed framework is illustrated in Fig. 1.

In order to evaluate the quality of visual themes, we ask human evaluators to examine how well these themes are in the task of describing similar visual contents. The result is pretty promising and demonstrate the effectiveness and rationality of discovered visual themes. We also explore potential applications of discovered VTs by designing three common CV tasks: example based image search, keyword based image search and image labelling. We work on four popular benchmarks, namely, Corel5K [1], NUS-Wide-Lite [2], IAPR-TC12 [3] and a subset of ESP-game [4]. The first two are used for example based search and keyword based search respectively. The last two and Corel5K are chosen as the testbeds of image labelling. We show the usefulness and advantages of using VTs rather than individual tags for these tasks.

## 2 Related Work

Our definition of visual theme is partly inspired by the naming of visual concept [5]. A visual concept is denoted as a subset of human language vocabularies that refer to particular visual entities (e.g. fireman, policeman). Visual concepts have long been collected and used by computer vision researchers in multiple domains [6][7][8][9]. A example in image analysis is ImageNet [10], where visual concepts (only nouns) are selected and organised hierarchically on the basis of WordNet [11]. A drawback of visual concepts is, they are often manually defined, and sometimes they may fail to capture complex information within the visual world. This makes them less applicable in multiple domains.

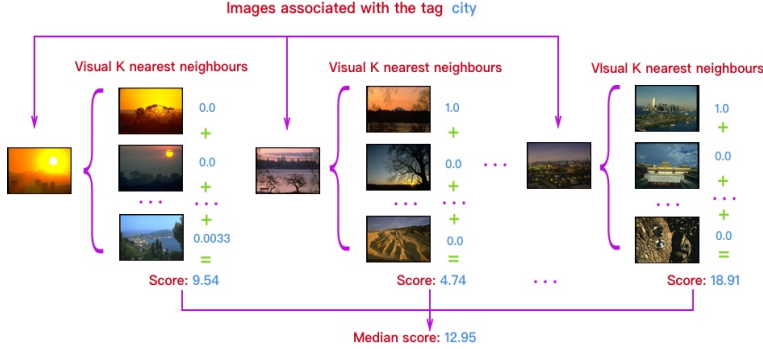
The subjectivity of visual concept definition hinders its extension to be used on different joint image and text databases. This motivates us to explore objective visual theme directly from raw images and associated tags. Our work on visual theme discovery is related to previous work on concept discovery [5][12][13]. In particular, LEVAN [13] starts with a given groups of general concepts and gradually divide them into subconcepts according to massive resources of online books. VisKE [12] focuses on validating relationship between pairs of concept entities from semantic and visual aspects. [5] builds a large amount of classifiers for terms filtering and similarity computation, then cluster selected terms into concepts.

A significant difference between our work and previous work is that we are not trying to build large amount of general visual concepts so as to describe as many image as possible, instead, we put forward an unsupervised and efficient framework to allow different image databases to have their own collection of visual themes as visual content description. Considering the quantity and diversity

of images, dividing large image collections into visual theme based categories can facilitate various tasks such as management, indexing and retrieval.

### 3 Visual Theme Discovery

This section elaborates the theme discovery workflow. Recall that a visual theme is constructed by a subset of tags which are capable of representing similar visual contents. To make it practical, we argue a VT should show strong connection to certain visual content that can be easily processed by computer vision algorithms. Besides, tags (including synonyms) describing same or similar visual content should be grouped into the same theme in order to maintain compactness. Start with the image corpus and associated tags, we first pick tags which show high-level visual content descriptive power, then cluster them into a set of VTs based on visual and semantic similarity.



**Fig. 2.** Workflow of Weight K-Nearest Measure. Given a tag and its associated images, for each image, we find its visual K-nearest neighbours and examine if other images under the same tag frequently appear in the K neighbours. We compute a score (higher is better) of each associated image of given tag, then take the median to quantify the tag’s ability towards visual content description.

#### 3.1 Tag Filtering

As we mentioned before, not all tags show strong connection with visual contents. Before discovering the themes, we need to examine each tag’s ability of visual content description, and filter out ones who are not qualified. The idea to achieve this is simple: if a tag is good at describing particular visual content, the majority of its associated images should also share similar visual contents. Hence, the visual similarities between images under a tag can reflect the tag’s ability towards visual content description. As for implementation, we represent images using feature activations from the pre-trained convolution neural network

(CNN) model due to its excellent performance in content-based image retrieval [14]. We then define Weighted K-Nearest Measure (WKNM) as measurement of tag’s ability towards visual content description.

The procedure of WKNM is illustrated in Fig. 2. Given tag  $t_i$  and its associated image set  $F_i = \text{set}\{f_{i1}, f_{i2}, \dots, f_{ij}, \dots, f_{in}\}$ , for every related image  $f_{ij}$ ,  $K$  nearest neighbours based on the cosine distance of their visual features are obtained using cosine distance. Thus, the similarity score between image  $f_{ij}$  and other images in  $F_i$  could be computed as:

$$\text{Sim}(f_{ij}, F_i) = \sum_{k=1}^K \left(1 - \frac{k-1}{K}\right) \delta(t_i, f_{ijk}) \quad (1)$$

where  $\delta(t_i, f_{ijk})$  is an indicator function which equals to 1 if image  $f_{ijk}$  contains tag  $t_i$ , otherwise it is set to 0.  $K$  is the number of nearest neighbours of image  $f_{ij}$ . It could be noticed that,  $\delta(t_i, f_{ijk})$  is penalised by multiplying a weight according to the sequence in  $K$  neighbours (a closer neighbour has a smaller sequence index). Hence,  $\text{Sim}(f_{ij}, F_i)$  quantifies tag  $t_i$ ’s ability towards visual content description based on image  $f_{ij}$ .

We successively compute all similarity scores based on each image in tag  $t_i$ ’s associated image set  $F_i$ , then take the median score to quantify tag  $t_i$ ’s ability towards visual content description. We call such a median the Visual Content Descriptive Level (VCDL) of a given tag. A larger VCDL of a tag indicates it is good at describing certain visual contents. We choose the median because it is a robust statistic, even if dataset is biased, the median is unlikely to offer an arbitrarily large or small result. We repeat this procedure on all tags and eliminate those whose VCDLs fall below a certain threshold. Note that we do not need to examine each tag’s frequency of occurrence since the WKNM method has inherently done this.

**Table 1.** Example of filtered tags on Corel5K dataset.

Filtered tags	Evaluation
{f-16, kauai, oahu}	too specific
{whited-tailed, close-up}	too abstract
{art, festival}	too generic

Table 1 gives a few examples of filtered tags on Corel5K dataset. We could clearly see our method is able to automatically remove tags that are not suitable for visual theme discovery. However, when we take a look at the remaining tags, we found some of them are synonyms e.g. *jet* and *plane*. It is necessary to group them together since they are likely to confuse CV algorithms and introduce extra computational cost. Moreover, we notice some tags are often used together to describe particular visual content. For instance, in Corel5K dataset, *grizzly* only

appears together with *bears* in images containing bears. This motivates us to measure tag similarity both semantically and visually.

### 3.2 Tag Visual Similarity Measure

We measure tag visual similarity by examining their distance in the metric space. Suppose we have a visual space constructed by all images, and each image's distance could be evaluated by computing distances between their corresponding visual features. In this space, each tag could be represented by its associated images which are a subset of the whole image set in visual space. Hence, the well-known Hausdorff distance (HD) is quite appropriate to measure visual distance between two different tags.

The Hausdorff distance is defined as the maximum distance of a set to the nearest point in the other set [15]. In our case, the Hausdorff distance from tag  $A$  to tag  $B$  in visual space would be:

$$h(A, B) = \max_{a \in A} \{\min_{b \in B} \{dist(a, b)\}\} \quad (2)$$

where  $a$  and  $b$  are image feature based points of tags  $A$  and  $B$  in high-dimensional visual space,  $dist(a, b)$  is certain distance metric between these points. For simplicity, we take  $dist(a, b)$  as the Euclidian distance between  $a$  and  $b$ .

Since HD measures the relative position of points in visual space, it's more robust to position variations than other methods. However, HD method is quite sensitive to outliers, which makes it inappropriate to tackle noisy data. A modified version of HD is proposed in [16]:

$$h_{mod}(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \{dist(a, b)\} \quad (3)$$

where  $|A|$  is the number of images associated with tag  $A$ . A problem of this revised HD is it contains points whose pairwise distances are zero. Considering that an image's annotation often contains more than 1 tag, points  $a$  and  $b$  in  $dist(a, b)$  could refer to the same image. Hence, we revise the formula in (3) to remove this negative impact:

$$h'_{mod}(A, B) = \frac{1}{|A'|} \sum_{a \in A'} \min_{b \in B} \{dist(a, b)\} \quad (4)$$

where  $A' = \min_{b \in B} \{dist(a, b) \neq 0\}$ . We use (4) to measure visual distance from tag  $A$  and tag  $B$ . Since associated images of different tags also differ, we modify the final Hausdorff distance between two tags as:

$$F'(A, B) = \max\{h'_{mod}(A, B), h'_{mod}(B, A)\} \quad (5)$$

Ultimately we can obtain a distance matrix  $M_{vdist}$  where each entry is the visual distance between two tags. It's easy to switch distance to similarity: just rescale all values in  $M_{vdist}$  to the range from 0 to 1, then replace each entry

value with the difference between 1 and original value. We denoted the tag visual similarity matrix as  $M_{vsim}$ . Larger values in  $M_{vsim}$  indicates stronger visual similarity between two corresponding tags.

### 3.3 Tag Semantic Similarity Measure

We measure semantic similarity between two tags by evaluating their word embeddings [17] [18] in an unsupervised manner. In the embedding space, each distinct word is represented using a N-dimensional vector. The embedding algorithm first assign each word vector with random values, then recursively adjust the value of these vectors according to some objective function. More specifically, we train a Skip-gram neural network language model [17] on latest dump of English Wikipedia using Word2Vec [19] toolset.

To elaborate, the training set is a large collection of English Wikipedia articles. In the training phase, each time a short sequence of words are extracted from an article using a sliding window with fixed width. Then the corresponding word vectors (random values at first) are extracted and fed into the skip-gram model. The training objective is to enable words to effectively predict nearby words, so words enjoy higher semantic similarity lie closer in the semantic space.

Once training process is completed, we extract word vectors from the trained model according to the content of tags, then evaluate semantic similarity of each pair of tags by computing cosine distance between their corresponding word vectors. Similarly, we build the the semantic similarity matrix  $M_{ssim}$ . Again, we replace each entry value in  $M_{ssim}$  with the difference between 1 and original value. Larger values in  $M_{ssim}$  indicates stronger semantic similarity between two corresponding tags.

### 3.4 Clustering Tags into Visual Themes

With two similarity matrices  $M_{vsim}$  and  $M_{ssim}$ , we linearly merge them into joint similarity matrix  $M_{join}$  via a parameter  $\alpha$  (from 0 to 1). We can control the proportion of visual and semantic components by tuning  $\alpha$ .

$$M_{join} = \alpha \times M_{vsim} + (1 - \alpha) \times M_{ssim} \quad (6)$$

Based on  $M_{join}$ , we use spectral clustering [20] to cluster tags into a collection of visual themes. Table 2 describes a few themes discovered on Corel5K dataset with  $\alpha$  fixed to 0.12. Note that although  $\alpha$  could be set to 0, which means no visual clue is used for similarity measurement, however, that might lead to sub-optimal result since semantic similarity mainly depends on word co-occurrence in text corpus.

## 4 Human Evaluation of Visual Themes

After clustering phase, each visual theme is represented as a set of tags and associated images. As we mentioned in Section 3, a visual theme should show

**Table 2.** Example of visual themes discovered on Corel5K dataset.

Concept Type	Concept Content
scene	{sunrise, sunset}
object	{mare, foals, horses}
mixed	{cloud, sky, mist, horizon}
mixed	{jet, flight, runway, plane}

strong connection to certain visual content. Besides, tags (including synonyms) describing the same or similar visual content should be grouped into the same theme. Hence, we design a human evaluation experiment to examine quality of discovered visual themes from these two aspects.

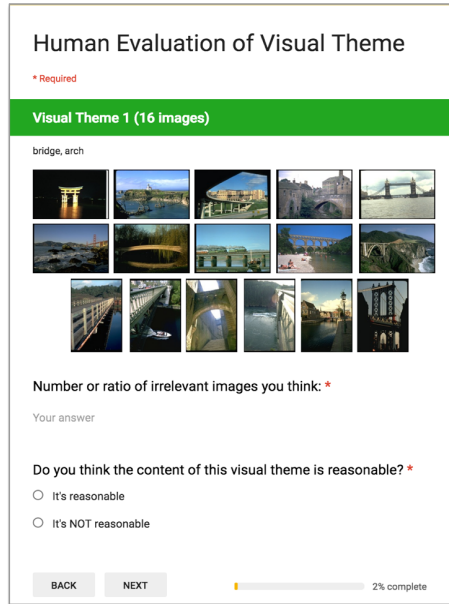
**Evaluation setup:** We use Corel5K dataset and discover 100 visual themes using 4500 training images and associated tags. We feed training images into the VGG-16 [21] model and take the output of fully-connected layer 'fc7' (4,096 dimensions) as image-level holistic visual features. Then we choose 499 testing images as evaluation set, and replace tag based annotation with corresponding visual themes. Hence, testing images are categorised into a collection of visual themes. Next we remove themes whose frequencies of occurrence are less than 3 times across all testing images, and keep 66 visual themes for evaluation. A print version of evaluation examples and interface could be found in the supplementary materials.

We designed a two-step procedure for human evaluation. A example of the evaluation interface is shown in Fig. 3. For each visual theme, we first display its tags and associated images to human evaluators, then asked them to examine whether the visual content described by this visual theme appears in every associated images. If not, they need to give number of images which they think are relevant to the given theme. Thus we can easily compute the ratio of relevant images for each visual theme, and we name such a ratio as accuracy of visual content description (AVCD) of a visual theme. The AVCD for each visual theme is obtained by averaging all evaluators' responses on that theme.

In the following step we asked evaluators to examine tags contained in visual themes. They need to check if all tags within a visual theme are semantically connected and refer to similar visual content. If so, the corresponding visual theme is regarded as rational and vice versa. The final decision of rationality for each visual theme was combined using majority vote of human evaluators.

17 human subjects participated in the evaluation experiment and result is summarised in Fig. 4. In (a) we can clearly see that more than half of discovered visual themes achieve an accuracy over 0.9 on visual content description, and only 4% of them did not perform well on this task. In terms of rationality, 92% of visual themes are voted as rational while the remaining 8% are not. The experiment result demonstrates the effectiveness of discovered visual themes towards visual content description.





**Human Evaluation of Visual Theme**

\* Required

**Visual Theme 1 (16 images)**

bridge, arch

Number or ratio of irrelevant images you think: \*

Your answer

Do you think the content of this visual theme is reasonable? \*

☐ It's reasonable

☐ It's NOT reasonable

BACK NEXT 2% complete

**Fig. 3.** An example of human evaluation interface. Human evaluators need to give numbers of images which are irrelevant to current displayed visual theme, and also vote for the rationality of this visual theme.

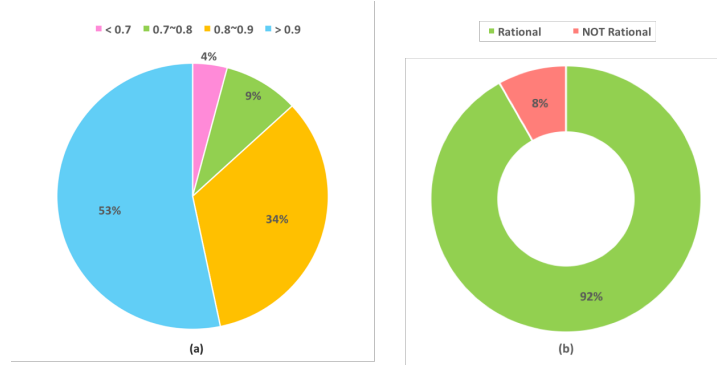
## 5 Application of Visual Themes

After human evaluation of visual themes, we further show potential applications of visual themes via three common computer vision experiments: example based image search, keyword based image search and image labelling.

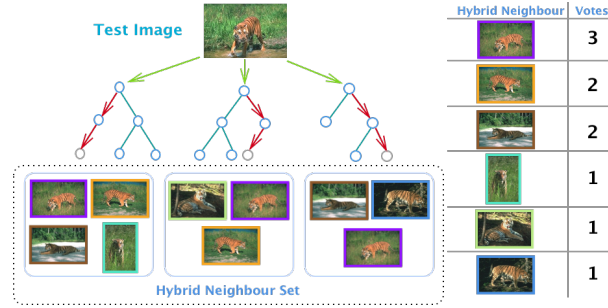
### 5.1 Construct an Image Retrieval and Labelling Framework

For building the framework, a vital issue need to be taken into consideration: how to design an effective data structure in terms of storage and speedy retrieval. Inspired by [22], we construct random forest using image features and discovered visual themes. In each random tree, we do binary split on visual features, and evaluate the split by computing histogram of visual themes. The well-know information gain[23] is used as the objective function.

The architecture of random forest is illustrated in Fig. 5. Given a test image, we feed its visual feature into one random tree, and it keeps falling until it reaches a leaf node. Consequently, training examples under the same leaf node share similar or same visual themes with the test image. Here we name a related training example as a Hybrid Neighbour (HN). We successively feed the test image to all random trees and obtain the Hybrid Neighbour Set (HNS) which is formed by all HNs. Additionally, the frequency of occurrence for a single HN



**Fig. 4.** Result of human evaluation of discovered visual themes on Corel5K dataset. (a): Result on accuracy of visual content description of visual themes. (b): Evaluators' responses on rationality of visual themes.



**Fig. 5.** Architecture of random forest for image retrieval and labelling. The visual feature of test image is put into the forest and similar images in training set will be found. Training images with higher frequency of occurrence will enjoy a higher rank in the returned result.

in HNS is defined as Hybrid Neighbour Vote (HNV). Apparently, a larger HNV indicates stronger similarity between a train image and the test image, and vice versa.

## 5.2 Example Based Image Search

**Scenario.** The retrieval system accept an image as input and then returns a list of ranked images according to some similarity measure. In our case, we just put the test image into the random forest and obtain its HNS and corresponding HNVs. The returned images are then ranked by their HNVs following an descending order. Usually the top K results will be returned by the retrieval system.

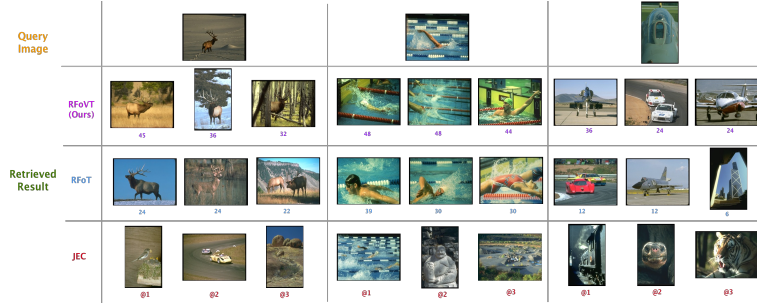
**Data.** We work on the popular Corel5K [1] benchmark which contains 4999 images. It is commonly split into 4500 image for training and the remaining 499 for testing, and 260 tags appear in both of these two sets.

**Evaluation metric.** Since Corel5K dataset does not have ground truth images for this task, we use K-Nearest Semantic Measure (KNSM) defined in [22] as evaluation metric:

$$KNSM = \sum_{q=1}^Q \sum_{t=1}^T \sum_{k=1}^K \delta(H_{qk}, t) \quad (7)$$

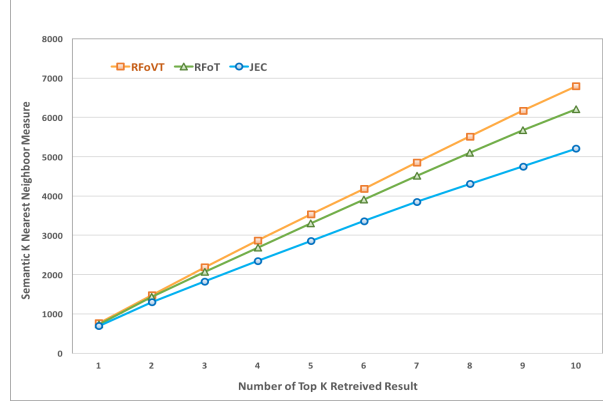
where  $Q$  is number of queries,  $T$  denotes the number of tags contained in query image and  $K$  represents that top  $K$  retrieved Hybrid Neighbours.  $\delta(H_{qk}, t) = 1$  if query image  $q$ 's tag  $t$  appears in its  $k^{th}$  HN, and  $\delta(H_{qk}, t) = 0$  if not. Hence a larger KNSM indicates stronger similarity between query image and its HNs since they share more tags.

**Parameter setting.** We eliminate tags whose visual content description levels (VCDLs) fall below 1.5, which results in 25 tags removed from original tag set. Next, 100 visual themes are obtained by clustering 235 remaining tags. Then we construct 400 random trees for image to image search. We also reproduce the result in [22] to justify the superiority of VSCC over tags. In terms of the baseline method, we select Joint Equal Contribution (JEC) [24] where various types of features are equally weighted for visual distance measurement, and is shown to perform well in image retrieval and annotation.



**Fig. 6.** Qualitative result of example based image search

**Result.** Fig. 6 shows some qualitative results of three methods: random forest on visual themes (RFoVT), random forest on tags (RFoT) [22] and JEC. Pink numbers under the result images denote their corresponding HNVs, the blue number is similar to HNV, but it's computed based on tags in stead of visual themes. Magenta numbers means the rankings of returned images using JEC method. Apparently RFoVT and RFoT greatly outperforms the JEC counterpart. Moreover, our RFoVT performs slight better than RFoT both in normal case (see first example) and hard cases (see last example).



**Fig. 7.** KNSM measure of example based image search.

We also provide quantitative analysis using KNSM. We perform retrieval using all 499 testing images and result is illustrated in Fig. 7. Clearly our method finds images with higher semantic similarity than the other two methods. Our success on this task demonstrates that visual themes are better than tags in terms of visual content description.

### 5.3 Keyword Based Image Search

**Scenario.** Given a query keyword, the retrieval system returns a collection of images that are most likely to contain that word. On this task, we tend to use a large image repository where training instances are annotated with tags while testing instances are not.

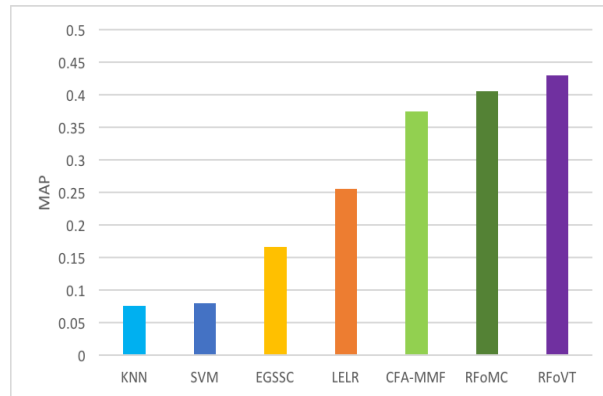
**Data.** We consider NUS-Wide-Lite dataset which contains 55,615 images, half of them (27,807) are used for training and the other half (27,808) for testing. We directly use 1,000 tags provided by the author for visual theme discovery. There are also 81 manually defined concepts available in dataset, each concept is represented with a single word.

**Parameter setting and evaluation metric.** We first remove tags whose VCDLs are below than 2.5, then cluster 904 remaining tags into 300 visual themes. We build 400 random trees and evaluate the proximity between a test instance  $T$  and a visual theme  $c$  as:

$$p(T, c) = \frac{\sum_{n=1}^N \delta(h_n, c) v_n}{\sum_{n=1}^N v_n} \quad (8)$$

where  $N$  is the size of hybrid neighbour set (HNV) of instance  $T$ ,  $h_n$  denotes a hybrid neighbour (HN) in HNV, and  $v_n$  denotes the hybrid neighbour votes (HNV) of  $h_n$ .  $\delta(h_n, c)$  is an indicator function which equals to 1 if visual theme  $c$  exists in  $h_n$ , and is equal to 0 otherwise.

In experiment, we treat each visual theme as a whole keyword, that means searching using any tags within same visual theme will obtain same results. We compare the Mean Average Precision (MAP) achieved on visual themes (RFoVT) with five previous methods on 81 manually defined concepts, namely, K Nearest Neighbour (KNN), Support Vector Machine (SVM) [25], Entropic Graph Semi-Supervised Classification (EGSSC) [26], Label Exclusive Linear Representation (LELR) [27], and Feature Analysis and Multi-Modality Fusion (CFA-MMF) [28]. Additionally we repeat the work in [22] and construct another random forest using 81 manually defined concepts (RFoMC), then perform the same task.



**Fig. 8.** MAP of keyword based image search on NUS-WIDE-Lite.

**Result.** The overall results are shown in Fig. 8. We can clearly see that some of previous methods have achieved much higher MAP than the KNN baseline on 81 manually selected concepts, but they still fail to achieve a MAP over 40%. While our random forest on visual themes (RFoVT) could obtain a MAP of 42.96%. This result demonstrates automatically discovered visual themes could do better than manually selected concepts in terms of visual content representation.

#### 5.4 Image Labelling

In order to further explore the potency of visual themes, we perform image labelling experiment on three well-known benchmarks: Corel5K [1], IAPR-TC12 [3] and a subset of ESP-game [4]. Table 3 provides details of three datasets and empirical settings of this task.

In this task, we do not perform tag filtering and only calculate the VCDLs for all tags. Given a test image, we put it into the random forest and obtain its HNs, and retain top voted  $m$  HNs according to their HNVs. Then we collect all tags within these HNs and keep at most  $n$  tags with highest VCDLs as final

**Table 3.** Details of three image datasets and experimental parameters.

Dataset	Corel5K	IAPR-TC12	ESP Game
Number of training samples	4500	17665	18689
Number of testing samples	499	1962	2081
Number of tags	260	291	268
$\alpha$ (Merging similarity matrix)	0.15	0.3	0.2
Number of random trees	400	400	400
Number of top voted HNs	3	3	3
Number of tags returned	up to 5	up to 5	up to 5

results. It’s a natural approach since selected tags are visually and semantically connected to the test image.

**Table 4.** Image annotation results on three datasets.

Dataset	Corel5K		IAPR-TC12		ESP Game	
Method	Precision	Recall	Precision	Recall	Precision	Recall
MBRM [29]	0.24	0.25	0.24	0.23	0.18	0.19
JEC [24]	0.27	0.32	0.28	0.29	0.22	0.25
TagProp [4]	0.33	0.42	0.46	0.35	0.39	0.27
GS [30]	0.30	0.33	0.32	0.29	-	-
SML+RF [31]	0.36	0.33	0.27	0.30	-	-
RF_optimize [32]	0.29	0.40	0.45	0.31	0.41	0.26
RFoVT	<b>0.40</b>	<b>0.35</b>	0.31	0.23	0.29	0.20

We report average precision and average recall of image labelling with comparison to previous works in Table 4. From the table we can see that our method (RFoVT) outperforms all previous methods on Corel5K dataset, but its performance falls behind TagProp [4] and RF\_optimize [32] on the other two datasets. However, the success of TagProp largely depends on its tedious optimisation for each image and tag, which hinders its extension to large scale dataset. While RF\_optimize treats each tag as an independent unit and ignore their visual and semantic connection, which makes it less competent in dealing with noisy data. Note that web images in real world often come with considerable amount of redundant and unnecessary information. On the contrary, our image labelling method can be easily extended to large scale dataset, and can easily eliminate the majority of noisy dataset by applying tag filtering procedure. Although RFoVT does not perform very well on all datasets, it is quite simple yet efficient considering the intrinsic architecture of random forest. The result may be improved by adopting more sophisticated tag selection algorithm.

## 6 Concluding Remarks

In this paper, we put forward an unsupervised framework to automatically discover visual theme which can effectively describe visual contents. Then we perform manual evaluation to evaluate the quality of discovered visual themes, and show their potential applications in computer vision field via three common tasks. The results of example based image search, keyword based image search and image labelling experiments demonstrate the effectiveness and versatility of discovered visual themes.

## References

1. Duygulu, P., Barnard, K., de Freitas, J.F., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Computer Vision/ECCV 2002*, Springer (2002) 97–112
2. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: *Proceedings of the ACM international conference on image and video retrieval*, ACM (2009) 48
3. Hugo Jair Escalante, Carlos A Hernandez, J.A.G.A.L.M.M.E.F.M.L.E.S.L.V.M.G.: The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding* **114** (2010)
4. Matthieu Guillaumin, Thomas Mensink, J.V.C.S.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. (2009)
5. Sun, C., Gan, C., Nevatia, R.: Automatic concept discovery from parallel text and visual corpora. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2015) 2596–2604
6. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE (2012) 1234–1241
7. Zhou, B., Jagadeesh, V., Piramuthu, R.: Conceptlearner: Discovering visual concepts from weakly labeled image collections. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2015) 1492–1500
8. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2015) 1473–1482
9. Singh, B., Han, X., Wu, Z., Morariu, V.I., Davis, L.S.: Selecting relevant web trained concepts for automated event retrieval. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2015) 4561–4569
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE (2009) 248–255
11. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38** (1995) 39–41
12. Sadeghi, F., Divvala, S.K., Farhadi, A.: Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, IEEE (2015) 1456–1464

13. Divvala, S., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3270–3277
14. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Computer Vision–ECCV 2014. Springer (2014) 584–599
15. Shonkwiler, R.: Computing the hausdorff set distance in linear time for any lp point distance. *Information Processing Letters* **38** (1991) 201–207
16. Dubuisson, M.P., Jain, A.K.: A modified hausdorff distance for object matching. In: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. Volume 1., IEEE (1994) 566–568
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119
18. Bengio, Y., Schwenk, H., Sen  cal, J.S., Morin, F., Gauvain, J.L.: Neural probabilistic language models. In: Innovations in Machine Learning. Springer (2006) 137–186
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: word2vec (2014)
20. Von Luxburg, U.: A tutorial on spectral clustering. *Statistics and computing* **17** (2007) 395–416
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Fu, H., Qiu, G.: Fast semantic image retrieval based on random forest. In: Proceedings of the 20th ACM international conference on Multimedia, ACM (2012) 909–912
23. Chen, X., Mu, Y., Yan, S., Chua, T.S.: Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In: Proceedings of the international conference on Multimedia, ACM (2010) 35–44
24. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. *International Journal of Computer Vision* **90** (2010) 88–105
25. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann (2005)
26. Subramanya, A., Bilmes, J.A.: Entropic graph regularization in non-parametric semi-supervised classification. In: Advances in Neural Information Processing Systems. (2009) 1803–1811
27. Chen, X., Yuan, X.T., Chen, Q., Yan, S., Chua, T.S.: Multi-label visual classification with label exclusive context. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 834–841
28. Ha, H.Y., Yang, Y., Fleites, F.C., Chen, S.C.: Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval. In: Multimedia and Expo (ICME), 2013 IEEE International Conference on, IEEE (2013) 1–6
29. S L Feng, R Manmatha, V.L.: Multiple bernoulli relevance models for image and video annotation. (2004)
30. Shaoting Zhang, Junzhou Huang, Y.H.Y.H.L.D.M.: Automatic image annotation using group sparsity. (2010)
31. Motofumi Fukui, Noriji Kato, W.Q.F.X.: Multi-class labeling improved by random forest for automatic image annotation. (2011)
32. Hao Fu, Qian Zhang, G.Q.: Random forest for image annotation. *European Conference on Computer Vision* (2012)